5-17-2006

Determination of Amino Acid Composition of Soybeans (Glycine max) by Near-Infrared Spectroscopy

Igor V. Kovalenko
Iowa State University

Glen R. Rippke
Iowa State University

Charles R. Hurburgh Jr.
Iowa State University, tatry@iastate.edu

JOURNAL OF AGRICULTURAL AND FOOD CHEMISTRY

Determination of Amino Acid Composition of Soybeans (*Glycine max*) by Near-Infrared Spectroscopy

IGOR V. KOVALENKO,* GLEN R. RIPPKE, AND CHARLES R. HURBURGH

Department of Agricultural and Biosystems Engineering, Iowa State University, Ames, Iowa 50011

Calibration equations for the estimation of amino acid composition in whole soybeans were developed using partial least squares (PLS), artificial neural networks (ANN), and support vector machines (SVM) regression methods for five models of near-infrared (NIR) spectrometers. The effects of amino acid/protein correlation, calibration method, and type of spectrometer on predictive ability of the equations were analyzed. Validation of prediction models resulted in r^2 values from 0.04 (tryptophan) to 0.91 (leucine and lysine). Most of the models were usable for research purposes and sample screening. Concentrations of cysteine and tryptophan had no useful correlation with spectral information. Predictive ability of calibrations was dependent on the respective amino acid correlations to reference protein. Calibration samples with nontypical amino acid profiles relative to protein would be needed to overcome this limitation. The performance of PLS and SVM was significantly better than that of ANN. Choice of preferred modeling method was spectrometer-dependent.

KEYWORDS: Near-infrared (NIR) spectroscopy; soybeans; *Glycine max*; amino acids; chemometrics; partial least squares (PLS); artificial neural networks (ANN); support vector machines (SVM)

INTRODUCTION

Soybeans are a main source of plant protein for animal feed formulation. Modern diet formulation methods balance rations on the basis of amino acid content. This has increased the need for the development of rapid and cost-effective techniques for amino acid measurement.

Amino acid composition is normally determined using HPLC (*1*). This method is too slow and expensive for feed formulation and plant-breeding applications, when large numbers of samples have to be screened. Near-infrared (NIR) spectroscopy has been applied to amino acid analysis by several researchers with various degrees of success.

In feed formulation research, Irish et al. (*2*) provided an example of the application of NIR calibrations for comparison of protein, lysine, and total sulfur amino acid content of raw materials from different suppliers. The authors showed that this rapid analysis allowed for more efficient ration formulation through the detection and reduction of variation in feed ingredient composition. Van Kempen and Simmins (*3*) evaluated NIR technology for the estimation of digestible amino acid content in several feed ingredients of animal origin. Cross-validation of their calibration models for the prediction of lysine and methionine resulted in determination coefficient (r^2) ranging from 0.80 to 0.95.

In grain-related research, Williams et al. (*4*) reported satisfactory results ($r^2 = 0.66-0.96$) in correlating NIR spectral data of ground wheat and barley to their amino acid concentrations. Wu et al. (*5*) showed the applicability of NIR spectroscopy for the amino acid analysis of milled rice powder. In their study, most of the amino acid calibration models had high determination coefficients ($r^2 = 0.85-0.98$), except for those of cysteine ($r^2 = 0.78$), histidine ($r^2 = 0.65$), and methionine ($r^2 = 0.10$). An experiment conducted by Pazdernik et al. (*6*) demonstrated that the accuracy of NIR screening for amino and fatty acid concentrations in soybeans may be improved by grinding seed samples. In a large study of amino acid profiling in ground grain samples and various feed ingredients done by Fontaine et al. (*7*, *8*), r^2 values of 0.84−0.98 were obtained for soybeans and soybean meal. Overall, the predictive ability of amino acid calibration models was dependent on the type of grain, sample form (whole grain or ground), and specific amino acid. In addition, as suggested by Baianu et al. (*9*), the accuracy of NIR measurement of amino acids in soybeans may be affected by close correlation between some amino acids and total protein content. However, little is known on the effect of calibration (regression) method and the type of NIR spectrometer. The effect of correlation of amino acid concentration with protein has also not been studied in detail. Therefore, the objectives of this experiment were to (i) develop calibration equations for estimation of amino acid composition in whole soybeans using three linear and nonlinear regression methods for five models of NIR spectrometers and (ii) analyze the effects of amino acid/protein correlation, calibration method, and type of spectrometer on the predictive ability of the equations.

* Address correspondence to this author at the Department of Agricultural and Biosystems Engineering, 1551 Food Sciences Building, Iowas State University, Ames, IA 50011 [telephone (515) 294-6358; fax (515) 294-6383; e-mail igorkov@iastate.edu].

10.1021/jf052570u CCC: $33.50
Published on Web 04/20/2006

Table 1. Specifications of the NIR Spectrometers

	instrument				
characteristic	FOSS Infratec 1241 Grain Analyzer	DICKEY-john OmegAnalyzerG	Perten DA 7200	Bruker Optics/ Cognis QTA	ASD LabSpec Pro
technology	scanning monochromator, Si detector	scanning monochromator, Si detector	InGaAs photodiode array	FT-NIR, RT-PbS detector	Si and InGaAs photodiode arrays
mode	transmittance	transmittance	reflectance	reflectance	reflectance
spectral range	850–1048 nm	730–1100 nm	950–1650 nm	12000–4000 cm^{-1} (833–2500 nm)	350–2500 nm
spectral resolution	7 nm	not available	3.125 nm/diode	2–256 cm^{-1} [a]	3 nm, 10 nm
sampling interval	2.0 nm	0.5 nm	5.0 nm	7.7 cm^{-1}	1.4 nm, 2.0 nm
no. of data points	100	741	141	1037	2151

[a] Spectral resolution of 16 cm^{-1} was recommended by the manufacturer for this study.

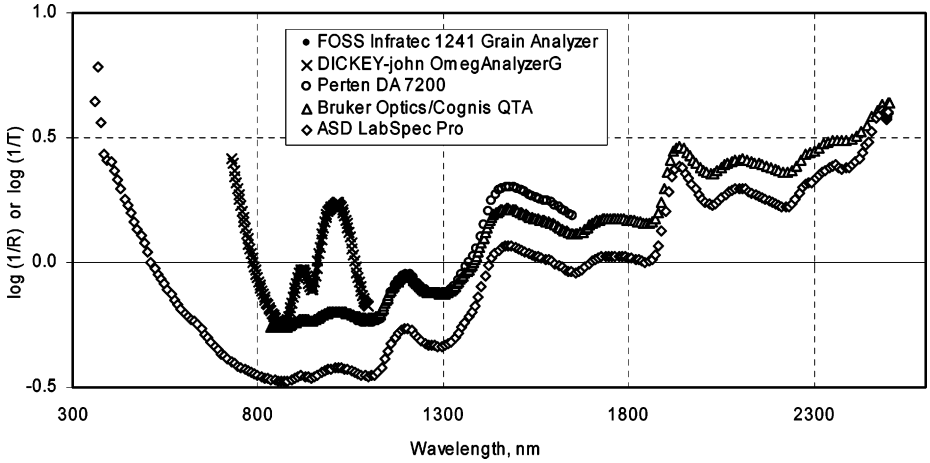

Figure 1. Normalized NIR scans of the same whole soybean sample obtained with five spectrometers.

MATERIALS AND METHODS

Raw Data. A calibration set of 526 soybean samples from the 1997–2001 crops and a test set of 147 samples (both sets contained various lines from all regions of the United States) from the 2002 crop were used for model development and testing. NIR spectra of the whole soybeans were obtained from five NIR spectrometers: FOSS Infratec 1241 Grain Analyzer (FOSS North America, Eden Prairie, MN); DICKEY-john OmegAnalyzerG (DICKEY-john Corp., Auburn, IL), Perten DA 7200 (Perten Instruments Inc., Springfield, IL), Bruker Optics/Cognis QTA (Bruker Optics Inc., Billerica, MA, and Cognis Corp., Cincinnati, OH), and ASD LabSpec Pro (Analytical Spectral Devices Inc., Boulder, CO). Specifications of the instruments are provided in **Table 1**. **Figure 1** illustrates NIR scans of the same soybean sample obtained with the five spectrometers.

The following 18 amino acids were considered in this study: alanine (Ala), arginine (Arg), aspartic acid (Asp), cysteine (Cys), glutamic acid (Glu), glycine (Gly), histidine (His), isoleucine (Ile), leucine (Leu), lysine (Lys), methionine (Met), phenylalanine (Phe), proline (Pro), serine (Ser), threonine (Thr), tryptophan (Trp), tyrosine (Tyr), and valine (Val). Their concentrations were determined at the Experiment Station Chemical Laboratories, University of Missouri, using official method AOAC 982.30 E (a,b,c) Ch. 45.3.05 (*1*). The crude protein content of the soybean samples was measured at Eurofins US laboratory (Des Moines, IA) using method AOCS Ba 4e-93 (*10*). Statistics of reference amino acid and protein concentrations, including standard error of laboratory (SEL), are given in **Table 2**. SEL was calculated as

$$\mathrm{SEL} = \sqrt{\frac{\sum_{i=1}^{N}[\sum_{j=1}^{R}(y_{ij} - \bar{y}_j)^2/(R-1)]}{N}}, \qquad (1)$$

Table 2. Statistics of Reference Amino Acid and Protein Concentrations (Percentage of Total Weight on Dry Basis) in Calibration and Test Soybean Samples Used in This Experiment

constituent	r^2 with crude protein	min, % DB	mean, % DB	max, % DB	SD	SEL[a]
Ala	0.82	1.46	1.79	2.13	0.128	0.015
Arg	0.87	2.21	3.17	4.44	0.397	0.017
Asp	0.91	3.59	4.79	6.03	0.470	0.031
Cys	0.37	0.52	0.70	0.86	0.063	0.013
Glu	0.83	5.36	7.66	10.18	0.868	0.097
Gly	0.88	1.38	1.77	2.15	0.143	0.013
His	0.82	0.91	1.15	1.41	0.096	0.013
Ile	0.76	1.47	1.94	2.36	0.172	0.022
Leu	0.90	2.47	3.26	3.95	0.274	0.028
Lys	0.87	2.15	2.69	3.28	0.200	0.018
Met	0.53	0.48	0.61	0.76	0.048	0.010
Phe	0.88	1.54	2.16	2.68	0.207	0.015
Pro	0.73	1.46	2.04	2.65	0.225	0.024
Ser	0.60	1.43	1.92	2.58	0.209	0.039
Thr	0.75	1.29	1.62	1.96	0.117	0.010
Trp	0.20	0.32	0.50	0.66	0.064	0.032
Tyr	0.82	1.18	1.53	1.83	0.129	0.015
Val	0.73	1.51	2.06	2.54	0.186	0.027
crude protein	1.00	33.82	43.16	54.61	3.960	0.337

[a] Standard error of laboratory; for SEL calculation details, refer to Materials and Methods.

where y_{ij} is the jth replicate of the ith sample, $\bar{y}_j$ is the reference method mean value of all the replicates of the ith sample, N is the number of samples, and R is the number of replicates. Only those samples that had obviously erroneous spectra and/or concentration values were considered to be gross outliers to be excluded from calibration and test sets.

Multivariate Modeling. Three regression methods, partial least squares (PLS), artificial neural networks (ANN), and support vector machines (SVM), were used in this work.

PLS. PLS_Toolbox 3.0 (Eigenvector Research Inc., www.eigenvector.com) for MATLAB (The MathWorks Inc., www.mathworks.com) was used for PLS modeling. The number of latent variables was selected using five-block cross-validation on the training set. For a detailed description of PLS regression refer to Næs et al. (*11*).

ANN. Neural Network Toolbox for MATLAB (The MathWorks Inc., www.mathworks.com) was used for development of ANN calibration models. Feed-forward back-propagation networks were trained on 80% of the available training samples. The other 20% of the training set was utilized as an early stopping set to prevent overfitting during training process. Input dimensionality was reduced by PCA. The best number of network inputs (principal components) and number of neurons in one hidden layer was determined by five-block cross-validation on the training set. A tangent sigmoid function and linear function were used as activation functions of hidden layer neurons and an output neuron, respectively. For more details on the ANN method refer to Haykin (*12*), Cherkassky and Mulier (*13*), Borggaard (*14*).

SVM. Least-squares implementation of SVM algorithm (LS-SVM) and LS-SVMlab1.5 toolbox for MATLAB developed by Suykens et al. (*15*) were utilized for this part of the experiment. The radial basis function (RBF)

$$K(\boldsymbol{x}, \boldsymbol{x}_k) = \exp(-||\boldsymbol{x} - \boldsymbol{x}_k||^2/\sigma^2), \quad (2)$$

where σ^2 is the RBF bandwidth, was used as a kernel function. The best pair of complexity regularization parameter (required for model training) and RBF bandwidth for every amino acid was determined by five-block cross-validation on the training set. More information on SVM may be found in Suykens et al. (*15*), Vapnik et al. (*16*), Smola and Scholkopf (*17*), Cogdill and Dardenne (*18*).

Data Preprocessing. Two methods for reduction of the light scatter effect, multiplicative scatter correction (MSC) and differentiation using Savitzky−Golay algorithm (*19*), were considered as primary pretreatments for spectral data. On the basis of preliminary results, we expected differentiation to be superior due to its universal applicability to all calibrations, regardless of regression method and spectrometer. Optimal combination of Savitzky−Golay algorithm parameters—window size, polynomial order, and derivative order (first or second)—was established on the basis of standard error of cross-validation of PLS calibrations. Because determining optimal parameter sets for all three regression methods was not feasible due to the amount of computation time required, PLS optimization was done at the potential expense of limiting the other calibration methods relative to PLS regression. The best results were obtained with second derivative for all spectra except for DICKEY-john data, which required only first-order derivation. This was most likely because raw spectra from this instrument had already been corrected for baseline shift by the instrument software.

In addition to differentiation, spectral data (wavelengths) from all instruments were normalized to have zero mean and unity standard deviation. For more details on data transformations for each spectrometer refer to **Table 3**.

Univariate Regression against Crude Protein. As can be seen from **Table 2**, most of the amino acids are strongly correlated with crude protein. This indicates that most of the amino acid concentrations could be derived from known reference protein values. To assess the accuracy of this prediction method (and to compare it with NIR calibrations), linear univariate regression equations for every amino acid were developed and tested using samples from NIR calibration and validation data sets.

Model Validation. An independent test set of 147 samples was applied to all calibration models, and the following parameters characterizing their predictive ability were computed: (i) coefficient of determination, r^2; (ii) standard error of prediction corrected for bias, SEP; (iii) bias or mean difference between NIR-predicted and reference concentrations, d; and (iv) relative predictive determinant, RPD. Definitions of these parameters can be found in Williams and Norris (*20*) and in the AACC NIR calibration guideline 39-00 (*21*).

Table 3. Transformations (in Sequential Steps) Applied to Absorbance Data from Five Spectrometers

instrument	spectral data preprocessing
FOSS Infratec 1241 Grain Analyzer	(1) second derivative (5, 3)[a] (2) normalization
DICKEY-john Omeg-AnalyzerG	(1) first derivative (17, 2) (2) normalization
Perten DA 7200	(1) second derivative (5, 3) (2) normalization
Bruker Optics/ Cognis QTA	(1) delete noisy data points in the range of 12000–11533 cm^{-1} (833–867 nm) (2) smooth (37, 2) noisy spectra in the range of 11533–8910 cm^{-1} (867–1122 nm) (3) second derivative (25, 3) (4) normalization
ASD LabSpec Pro	(1) delete noisy data points in the range of 350–440 nm (2) use every other wavelength for subsequent steps (3) MSC (4) second derivative (9, 3) (5) normalization

[a] Parentheses contain window size and polynomial order for Savitzky–Golay differentiation algorithm.

RESULTS AND DISCUSSION

Overall Observations. Validation of the developed calibration models (PLS, ANN, and LS-SVM) resulted in coefficients of determination, r^2, ranging from 0.04 for Trp to 0.91 for Leu and Lys (in terms of RPD coefficients, predictive ability of the equations extended from 0.98 for Trp to 3.29 for Leu). On the basis of guidelines for interpretation of r^2 outlined by Williams and Norris (*20*), the division of NIR calibration models was as follows: (a) $r^2 = 0.00...0.25$, unusable models, Trp; (b) $r^2 = 0.26...0.49$, poor correlation models, Cys; (c) $r^2 = 0.50...0.64$, models usable for rough sample screening, Met and Se; (d) $r^2 = 0.66...0.81$, models usable for sample screening, Ala, Glu, Ile, Pro., Thr, and Val; (e) $r^2 = 0.83...0.90$, models "usable with caution for most applications", Arg, Asp, Gly, His, Leu, Lys, Phe, and Tyr. (Note: due to rounding off, there are no r^2 values of 0.65 and 0.82.)

Because of correlation between r^2 and RPD

$$\mathrm{RPD} = 1/(1 - r^2)^{0.5} \quad (3)$$

from personal communication with David B. Funk (*22*), the same classification of models could be derived from RPD values. Converting RPD into r^2 space and removing discontinuity due to rounding-off error give (a) RPD = 1.00...1.15, unusable models, Trp; (b) RPD = 1.16...1.40, poor correlation models, Cys; (c) RPD = 1.41...1.70, models usable for rough sample screening, Met and Ser; (d) RPD = 1.71...2.42, models usable for sample screening, Ala, Glu, Ile, Pro, Thr, and Val; and (e) RPD = 2.43...3.54, models "usable with caution for most applications", Arg, Asp, Gly, His, Leu, Lys, Phe, and Tyr.

The results of the experiment in terms of validation RPD and SEP of NIR calibration models (PLS, LS-SVM, and ANN) for five spectrometers are provided in **Tables 4−8**.

Attempts to explain the variation in models' predictive ability by correlating RPD of a specific calibration method to such properties of amino acids as average reference concentration in soybeans, relative variation of concentration (range divided by average), molecular weight, solubility in water, and isoelectric point did not result in any reliable relationship. However, when

Table 4. FOSS Infratec 1241 Grain Analyzer: Test Statistics (RPD and SEP) of the Three Types of Calibration Models

	PLS		LS-SVM		ANN	
amino acid	RPD	SEP	RPD	SEP	RPD	SEP
Ala	2.22	0.05	2.18	0.06	2.33	0.05
Arg	2.60	0.15	2.70	0.14	2.59	0.15
Asp	2.88	0.16	2.81	0.17	2.82	0.17
Cys	1.25	0.04	1.24	0.04	1.17	0.05
Glu	1.87	0.52	1.89	0.52	1.81	0.54
Gly	2.65	0.05	2.62	0.05	2.53	0.06
His	2.59	0.04	2.70	0.04	2.66	0.04
Ile	2.32	0.07	2.26	0.08	2.24	0.08
Leu	3.21	0.09	3.29	0.09	3.12	0.09
Lys	2.89	0.07	2.86	0.07	2.88	0.07
Met	1.50	0.03	1.54	0.03	1.44	0.03
Phe	2.82	0.07	2.82	0.07	2.80	0.07
Pro	2.28	0.09	2.27	0.10	2.30	0.09
Ser	1.44	0.17	1.43	0.17	1.34	0.18
Thr	2.01	0.06	1.98	0.06	1.94	0.06
Trp	1.02	0.08	0.98	0.08	0.99	0.08
Tyr	2.95	0.04	2.80	0.04	2.71	0.05
Val	2.07	0.08	1.97	0.08	2.05	0.08

Table 5. DICKEY-john OmegAnalyzerG: Test Statistics (RPD and SEP) of the Three Types of Calibration Models

	PLS		LS-SVM		ANN	
amino acid	RPD	SEP	RPD	SEP	RPD	SEP
Ala	2.25	0.05	2.24	0.05	2.26	0.05
Arg	2.05	0.19	2.74	0.14	2.58	0.15
Asp	2.62	0.18	2.93	0.16	2.73	0.17
Cys	1.19	0.05	0.98	0.06	1.14	0.05
Glu	1.76	0.56	1.84	0.54	1.73	0.57
Gly	2.48	0.06	2.37	0.06	2.41	0.06
His	2.69	0.04	2.93	0.04	2.59	0.04
Ile	2.24	0.08	2.16	0.08	2.14	0.08
Leu	3.02	0.09	3.23	0.09	3.03	0.09
Lys	2.68	0.08	2.96	0.07	2.79	0.07
Met	1.44	0.03	1.36	0.03	1.31	0.03
Phe	2.79	0.07	2.85	0.07	2.69	0.07
Pro	2.14	0.10	2.07	0.10	2.05	0.11
Ser	1.37	0.18	1.36	0.18	1.32	0.18
Thr	1.88	0.07	1.89	0.07	1.85	0.07
Trp	1.01	0.08	0.99	0.08	1.03	0.08
Tyr	2.70	0.05	2.79	0.04	2.67	0.05
Val	2.02	0.08	2.15	0.08	2.08	0.08

Table 6. Perten DA 7200: Test Statistics (RPD and SEP) of the Three Types of Calibration Models

	PLS		LS-SVM		ANN	
amino acid	RPD	SEP	RPD	SEP	RPD	SEP
Ala	2.16	0.06	2.11	0.06	2.13	0.06
Arg	2.89	0.14	2.27	0.17	2.61	0.15
Asp	2.93	0.17	2.82	0.17	2.81	0.17
Cys	1.25	0.04	1.17	0.05	1.22	0.04
Glu	1.84	0.55	1.83	0.56	1.84	0.56
Gly	2.65	0.05	2.56	0.06	2.27	0.06
His	2.87	0.04	2.73	0.04	2.51	0.04
Ile	2.28	0.08	2.10	0.08	2.00	0.09
Leu	3.24	0.09	3.14	0.09	2.78	0.11
Lys	2.69	0.08	2.56	0.08	2.60	0.08
Met	1.51	0.03	1.42	0.03	1.37	0.03
Phe	2.79	0.07	2.75	0.07	2.56	0.08
Pro	2.19	0.10	2.19	0.10	2.12	0.11
Ser	1.48	0.17	1.46	0.17	1.38	0.18
Thr	2.10	0.06	2.09	0.06	1.84	0.07
Trp	1.09	0.07	1.03	0.08	1.05	0.08
Tyr	2.64	0.05	2.63	0.05	2.67	0.05
Val	2.10	0.08	1.84	0.09	1.91	0.09

Table 7. Bruker Optics/Cognis QTA: Test Statistics (RPD and SEP) of the Three Types of Calibration Models

	PLS		LS-SVM		ANN	
amino acid	RPD	SEP	RPD	SEP	RPD	SEP
Ala	2.16	0.06	2.14	0.06	2.15	0.06
Arg	2.70	0.15	2.53	0.16	2.46	0.16
Asp	2.71	0.18	2.68	0.18	2.41	0.20
Cys	1.25	0.04	1.19	0.05	1.18	0.05
Glu	1.82	0.56	1.77	0.58	1.76	0.58
Gly	2.49	0.06	2.41	0.06	2.23	0.06
His	2.57	0.04	2.55	0.04	2.35	0.04
Ile	2.22	0.08	2.21	0.08	1.96	0.09
Leu	2.97	0.10	2.98	0.10	2.63	0.11
Lys	2.61	0.08	2.60	0.08	2.45	0.09
Met	1.45	0.03	1.34	0.03	1.40	0.03
Phe	2.62	0.08	2.90	0.07	2.60	0.08
Pro	2.17	0.10	2.15	0.10	2.05	0.11
Ser	1.43	0.17	1.36	0.18	1.25	0.20
Thr	1.87	0.07	1.82	0.07	1.70	0.07
Trp	1.07	0.07	0.98	0.08	1.01	0.08
Tyr	2.77	0.05	2.85	0.04	2.57	0.05
Val	1.97	0.08	2.04	0.08	2.03	0.08

NIR RPD values were regressed against determination coefficients describing the relationship between a particular amino acid with protein (**Table 2**), it became apparent that variation in NIR models' predictive ability was determined by how a certain amino acid was correlated to protein, as shown in **Figure 2**. If an amino acid concentration can be predicted from a known value of protein concentration, it can be estimated using NIR spectroscopy. If the correlation is poor, as is the case with Cys and Trp, NIR predictions will be equally inaccurate. A similar observation about the correlation between amino acid contents predicted by NIR and linear protein regression for soybean meal and full-fat soy was made by Fontaine et al. (*7*). This implies that NIR spectroscopy measures amino acid concentration in whole soybeans indirectly by deriving it from the total amount of nitrogen-containing molecules. Analysis of regression vectors of PLS calibration models supports this statement. Most of the regression vectors in **Figure 3** follow the same pattern, which indicates that, for the most part, the calibrations predicted protein. Those regression vectors that fall out of the general pattern represent low-RPD calibration models such as Trp, Cys, and Ser. Therefore, the biggest challenge that is faced in NIR measurement of amino acids in soybeans—and probably in other legumes and cereal grains—is to either exceed the correlation between amino acid and protein concentrations or assemble sample sets that break the correlation. Future research should attempt to address this issue by introducing calibration samples (possibly artificially created) with unusual amino acid profiles.

In general, r^2 values of this experiment were higher than those previously reported by Pazdernik et al. (*6*) for both whole-seed and ground-seed soybean samples. This could likely be attributed to a much larger calibration set used in this study (526 vs 90 samples) and form of expression of amino acid concentrations (percentage of total sample weight vs percentage of crude protein). Validation statistics of NIR calibrations for Leu, Lys, Met, and Thr in ground soybeans reported by Fontaine et al. (*7*) were superior to our results; however, these researchers reported higher correlations between protein and amino acids. Comparison of our results with those from previous studies by Pazdernik et al. (*6*), Fontaine et al. (*7*), Williams et al. (*4*), and Wu et al. (*5*) suggests that grinding grain samples may improve predictive ability of NIR spectroscopy in amino acid measurement. However, it is not clear whether grinding will make NIR predictions superior to the univariate protein regression results. Also, on the basis of the wheat and barley study by Williams

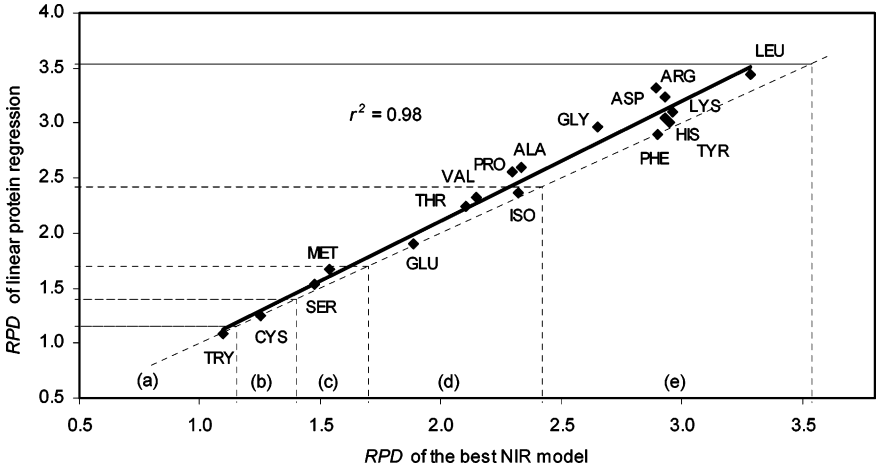

Figure 2. RPD of linear protein regression models versus RPD of the best NIR calibration models. Dotted lines define RPD regions described in the Overall Observations.

Table 8. ASD LabSpec Pro: Test Statistics (RPD and SEP) of the Three Types of Calibration Models

	PLS		LS-SVM		ANN	
amino acid	RPD	SEP	RPD	SEP	RPD	SEP
Ala	2.17	0.06	1.90	0.07	1.98	0.06
Arg	2.78	0.14	2.21	0.18	2.35	0.17
Asp	2.58	0.19	2.50	0.19	2.51	0.19
Cys	1.11	0.05	1.12	0.05	1.12	0.05
Glu	1.86	0.55	1.78	0.58	1.74	0.59
Gly	2.46	0.06	2.34	0.06	2.30	0.06
His	2.62	0.04	2.71	0.04	2.25	0.05
Ile	2.02	0.09	1.76	0.10	1.86	0.09
Leu	3.11	0.09	3.07	0.10	2.52	0.12
Lys	2.81	0.08	2.85	0.07	2.32	0.09
Met	1.49	0.03	1.42	0.03	1.35	0.03
Phe	2.79	0.07	2.74	0.07	2.30	0.09
Pro	2.21	0.10	2.19	0.10	2.02	0.11
Ser	1.42	0.17	1.32	0.19	1.35	0.18
Thr	1.90	0.07	1.72	0.07	1.75	0.07
Trp	1.06	0.07	1.02	0.08	0.99	0.08
Tyr	2.56	0.05	2.64	0.05	2.30	0.06
Val	1.98	0.08	1.67	0.10	1.84	0.09

et al. (*4*), additional improvement in amino acid NIR measurement may be gained by the selection of individual wavelengths (or regions) of NIR spectrum, as opposed to using full spectrum techniques.

As a side note, an interesting observation was made on the effect of spectral data preprocessing for one of the tested spectrometers. Although not normally done in practice, performing MSC with subsequent second-order differentiation improved RPD values of ASD LabSpec Pro calibrations up to 9% compared to differentiation alone. The advantage of additional preprocessing for this instrument might be explained by the more intensive light scatter in the far NIR range and the fact that this spectrometer uses more than one photodiode array detector.

Comparison of Multivariate Calibrations with Univariate Protein Regression. Validation results of linear protein regression models for calculation of amino acid concentrations are provided in **Table 9**. Comparison of overall performance of NIR calibrations to the protein regression was based on RPD coefficient, which is a standardized parameter of predictive ability. Analysis of ANOVA least squares model of the form

$$\mathrm{RPD} = \mathrm{AA} + M_1 + \mathrm{error} \qquad (4)$$

where AA is amino acid factor (18 levels) and M_1 is method factor (2 levels, protein regression and the best NIR calibration), demonstrated significance of the method factor ($p = 0.0002$). Comparison of means showed that RPD of protein regression was significantly higher than that of the best NIR calibration ($\alpha = 0.05$). As can be seen from **Tables 4**−**9**, in no case did the RPD of the NIR calibration (across three multivariate methods and five instruments) exceed the RPD of protein regression. NIR measurement of amino acid content is cumulative of two components: an error of protein prediction and an error of deriving amino acid content from predicted protein.

Comparison of Multivariate Calibration Methods and Spectrometers. Even though the prediction of amino acids was essentially a calculation from crude protein, comparison of modeling methods and spectrometers was still valid, in the sense that the results would apply to soybean measurement in general and could still indicate greater or lesser suitability for use. The effects of regression method and type of spectrometer on RPD was tested using ANOVA least-squares fit of the form

$$\mathrm{RPD} = \mathrm{AA} + M_2 + S + (M \times S) + \mathrm{error} \qquad (5)$$

where M_2 is method factor, S is spectrometer factor, and $M \times S$ is method−spectrometer interaction. The analysis indicated that all factors used in the model had a significant effect ($p < 0.0001$) on RPD. Due to a large number of samples and comparatively low sum of squares of $M \times S$, this interaction factor was ignored (included with the error term for comparison of means). As far as calibration methods were concerned, mean RPD values of calibration models (based on 18 amino acids times 5 spectrometers) were 2.19 (standard error of 0.063) for PLS, 2.16 (standard error of 0.066) for LS-SVM, and 2.08 (standard error of 0.059) for ANN. Means of PLS and LS-SVM regressions were not significantly different from each other, but were significantly higher ($\alpha = 0.05$) than mean RPD of ANN. The inferior performance of ANN could most likely be explained by either (i) an insufficient size of training set for this method or (ii) the use of PCA for dimensionality reduction of the input space, which discards information on nonlinearity that is contained in high-order principal components. Comparison of spectrometers demonstrated a significant advantage of the FOSS Infratec 1241 Grain Analyzer (**Table 10**, left column of means), despite its short optical range and small number of spectral data points. Additional data may be useful only if they add relatively more information than noise.

To determine whether the same calibration priority pattern, PLS−LS-SVM−ANN, applied to all spectrometers, levels of

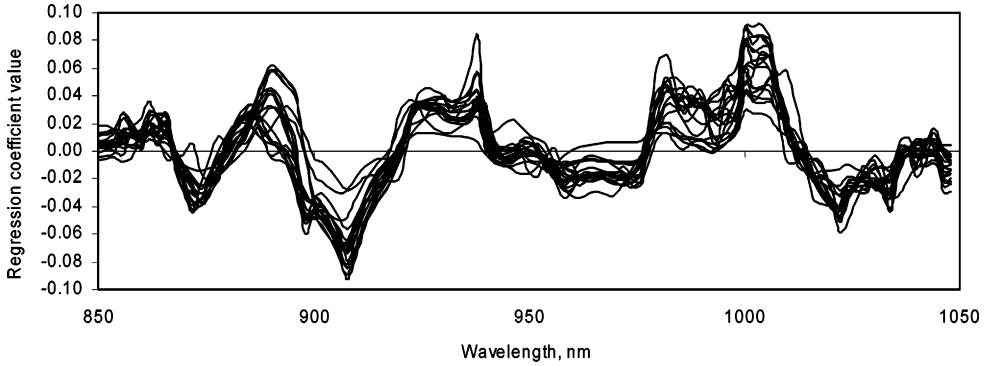

Figure 3. Regression vectors of 18 amino acid calibration models (PLS regression) developed for FOSS Infratec 1241.

Table 9. Test Statistics (RPD and SEP) of Linear Protein Regression Models

	protein regression	
amino acid	RPD	SEP
Ala	2.60	0.05
Arg	3.32	0.12
Asp	3.24	0.15
Cys	1.24	0.05
Glu	1.90	0.54
Gly	2.96	0.05
His	3.05	0.03
Ile	2.37	0.07
Leu	3.45	0.09
Lys	3.10	0.07
Met	1.67	0.02
Phe	2.90	0.07
Pro	2.55	0.09
Ser	1.53	0.16
Thr	2.24	0.06
Trp	1.09	0.07
Tyr	3.01	0.04
Val	2.32	0.07

Table 10. Mean RPD Values of Calibration Models for Five Spectrometers[a]

instrument	mean RPD based on 18 amino acids and 3 regression methods	mean RPD based on 18 amino acids and 1 best regression method
FOSS Infratec 1241 Grain Analyzer	2.23 a (0.085)[b]	2.25 ab (0.150)
Perten DA 7200	2.17 b (0.080)	2.26 a (0.148)
DICKEY-john Omeg-AnalyzerG	2.16 bc (0.086)	2.21 abc (0.166)
Bruker Optics/ Cognis QTA	2.10 cd (0.077)	2.16 c (0.135)
ASD LabSpec Pro	2.05 d (0.077)	2.163 bc (0.143)

[a] Means followed by the same lower case letter are not significantly different ($\alpha = 0.05$) by Tukey HSD test. [b] Parentheses contain standard error.

method−spectrometer interaction were analyzed (**Figure 4**). Results showed that the advantage of one calibration method over the others was spectrometer-dependent. Whereas the Perten DA 7200 and ASD LabSpec Pro had the largest differences in the mean RPD values across calibration methods, the FOSS Infratec 1241 Grain Analyzer showed nearly identical performance for all methods. The DICKEY-john OmegAnalyzerG, unlike the other spectrometers, demonstrated an advantage of the LS-SVM method over PLS.

Because RPD variations among calibration methods were not the same for all instruments, calibration models of the best-performing methods (PLS for FOSS Infratec 1241 Grain Analyzer, Perten DA 7200, Bruker Optics/Cognis QTA, and ASD LabSpec Pro; and LS-SVM for DICKEY-john OmegAnalyzerG) were used to further compare th performance of the spectrometers. Analysis of least-squares fit of the form

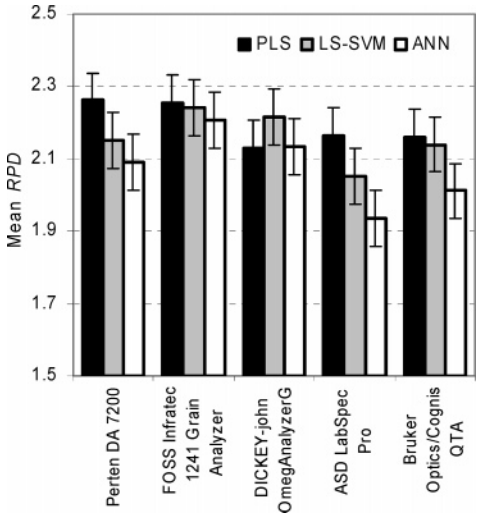

Figure 4. Mean RPD values (based on 18 amino acids) of calibration models grouped by spectrometers and calibration methods. Error bars indicate ± 3 standard errors.

$$\text{RPD} = \text{AA} + S + \text{Error} \quad (6)$$

demonstrated significance of amino acid (as expected) and spectrometer factors ($p < 0.0001$ and $p < 0.0029$, respectively). **Table 10** (right column of means) shows that the difference between spectrometers became less distinct when methods were pooled. The amino acid predictive ability of the Perten DA 7200, which had the highest mean RPD, was comparable to that of the FOSS Infratec 1241 Grain Analyzer and the DICKEY-john OmegAnalyzerG, but significantly better than those of the Bruker Optics/Cognis QTA and ASD LabSpec Pro.

An interesting observation was made by analyzing bias on the test set. Whereas average bias of all amino acid predictions from four spectrometers approached zero, all of the FOSS Infratec PLS calibration models except for the two unusable calibrations, Cys and Trp, had a negative bias, indicating that this spectrometer tended to overpredict amino acid concentrations. A completely opposite pattern was observed with this spectrometer in combination with LS-SVM and ANN calibration methods: all of the predictions except for Cys and Trp had a positive bias (figure not shown), indicating that with nonlinear calibrations the spectrometer underpredicted amino acid concentrations. This phenomenon could not be explained, because

the calibration and test sets were nearly identical for all spectrometers.

LITERATURE CITED

(1) AOAC. *Official Methods of Analysis of the Association of Official Analytical Chemists*, 15th ed.; AOAC International: Arlington, VA, 1990.

(2) Irish, G. G.; Fickler, J.; Fontaine, J. Practical application of near infrared reflectance spectroscopy to predict amino acids in feed ingredients. *Proc. Aust. Poult. Sci. Symp.* **2003**, *15*, 69.

(3) Van Kempen, T. A. T. G.; Simmins, P. H. Near-infrared reflectance spectroscopy in precision feed formulation. *J. Appl. Poult. Res.* **1997**, *6*, 471−477.

(4) Williams, P. C.; Preston, K. R.; Norris, K. H.; Starkey, P. M. Determination of amino acids in wheat and barley by near-infrared reflectance spectroscopy. *J. Food Sci.* **1984**, *49*, 17−20.

(5) Wu, J. G.; Shi, C.; Zhang, X. Estimating the amino acid composition in milled rice by near-infrared reflectance spectroscopy. *Field Crops Res.* **2002**, *75*, 1−7.

(6) Pazdernik, D. L.; Killam, A. S.; Orf, J. H. Analysis of amino and fatty acid composition in soybean seed using near infrared reflectance spectroscopy. *Agron. J.* **1997**, *89*, 679−685.

(7) Fontaine, J.; Horr, J.; Schirmer, B. Near-infrared reflectance spectroscopy enables the fast and accurate prediction of the essential amino acid contents in soy, rapeseed meal, sunflower meal, peas, fishmeal, meat meal products, and poultry meal. *J. Agric. Food Chem.* **2001**, *49*, 57−66.

(8) Fontaine, J.; Schirmer, B.; Horr, J. Near-infrared reflectance spectroscopy (NIRS) enables the fast and accurate prediction of the essential amino acid contents. 2. Results for wheat, barley, corn, triticale, wheat bran/middlings, rice bran, and sorghum. *J. Agric. Food Chem.* **2002**, *50*, 3902−3911.

(9) Baianu, I. C.; You, T.; Costescu, D. M.; Lozano, P. R.; Prisecaru, V.; Nelson, R. L. High-resolution nuclear magnetic resonance and near-infrared determination of soybean oil, protein, and amino acid residues in soybean seeds. In *Oil Extraction and Analysis: Critical Issues and Comparative Studies*; Luthria, D. L., Ed.; AOCS Press: Champaign, IL, 2004; pp 193−240.

(10) AOCS. *Official Methods and Recommended Practices of the American Oil Chemists' Society*, 5th ed.; AOCS Press: Champaign, IL, 1997.

(11) Næs, T.; Isaksson, T.; Fearn T.; Davies, T. *A User-Friendly Guide to Multivariate Calibration and Classification*; NIR Publications: Chichester, U.K., 2002.

(12) Haykin, S. *Neural Networks: A Comprehensive Foundation*, 2nd ed.; Prentice Hall: Englewood Cliffs, NJ, 1999.

(13) Cherkassky V.; Mulier, F. *Learning from Data: Concepts, Theory, and Methods*; Wiley: New York, 1998.

(14) Borggaard, C. Neural networks in near-infrared spectroscopy. In *Near-Infrared Technology in the Agricultural and Food Industries*, 2nd ed.; Williams, P., Norris, K., Eds.; AACC: St. Paul, MN, 2001; pp 101−107.

(15) Suykens, J. A. K.; Van Gestel, T.; De Brabanter, J.; De Moor, B.; Vandewalle, J. *Least Squares Support Vector Machines*; World Scientific: Singapore, 2002.

(16) Vapnik, V.; Golowich, S.; Smola, A. Support vector method for function approximation, regression estimation, and signal processing. In *Advances in Neural Information Processing Systems 9*; Mozer, M., Jordan M., Petsche, T., Eds.; MIT Press: Cambridge, MA, 1997; pp 281−287.

(17) Smola, A. J.; Scholkopf, B. *A Tutorial on Support Sector Regression*; NeuroCOLT2 Technical Report Series, NC-TR-98-030; Royal Holloway College, University of London: London, U.K., 1998.

(18) Cogdill, R. P.; Dardenne, P. Least-squares support vector machines for chemometrics: an introduction and evaluation. *J. Near Infrared Spectrosc.* **2004**, *12*, 93−100.

(19) Savitzky, A.; Golay, M. J. E. Smoothing and differentiation of data by simplified least-squares procedures. *Anal. Chem.* **1964**, *36*, 1627−1639.

(20) Williams, P., Norris, K., Eds. *Near-Infrared Technology in the Agricultural and Food Industries*, 2nd ed.; AACC: St. Paul, MN, 2001.

(21) AACC. *Approved Methods of the American Association of Cereal Chemists*, 10th ed.; AACC: St. Paul, MN, 2000.

(22) Personal communication with David B. Funk (Szent Istvan University, Budapest, Hungary) during the 12th International Diffuse Reflectance Conference (IDRC), Chambersburg, PA, 2004.

Received for review October 17, 2005. Revised manuscript received March 22, 2006. Accepted March 23, 2006. This journal paper of the Iowa Agriculture and Home Economics Experiment Station, Ames, IA, was supported by Hatch Act, State of Iowa, and United Soybean Board funds.

JF052570U